

Supporting information for;

Fitness costs of rifampicin-resistance in *Mycobacterium tuberculosis* are amplified under stringent conditions and compensated by mutation in the β' subunit of RNA polymerase.

Taeksun Song¹, Yumi Park¹, Isdore Chola Shamputa², Sunghwa Seo¹, Sun Young Lee¹, Han-Seung Jeon¹, Hongjo Choi¹, Myungsun Lee¹, Richard J. Glynne³, S. Whitney Barnes³, John R. Walker³, Serge Batalov³, Karina Yusim⁴, Shihai Feng⁴, Chang-Shung Tung⁴, James Theiler⁴, Laura E. Via², Helena I. M. Boshoff², Katsuhiko S. Murakami⁵, Bette Korber⁴, Clifton E. Barry, 3^{rd,1,2*}, Sang-Nae Cho^{1*}

¹ International Tuberculosis Research Center, Changwon, Republic of Korea,

² Tuberculosis Research Section, National Institute of Allergy and Infectious Disease, NIH, Bethesda, MD USA,

³ Genomics Institute of the Novartis Research Foundation, San Diego, CA USA,

⁴ Los Alamos National Laboratory, Los Alamos, NM USA,

⁵ Department of Biochemistry and Molecular Biology, The Center for RNA Molecular Biology, The Pennsylvania State University, University Park, PA USA.

*Correspondence to: Clifton E Barry, 3rd, email: cbarry@niaid.nih.gov or Sang-Nae Cho, email: raycho@yonsei.kr.

Table of Contents

Table S1. Disease characteristics, treatment regimens, and clinical outcome of subjects whose isolates were used in this study.

Table S2. Drug susceptibility patterns of isolates sequenced in this study.

Table S3. Sources of previously published full-length Mtb genomic sequences used in this study.

Table S4. Non-synonymous substitutions in 170 RIF-R clinical isolates.

Table S5. MIRU-VNTR patterns of selected isolates showing convergent acquisition of compensatory alleles and transmission of compensated strains.

Table S6. Bacterial strains and plasmids.

Figure S1. A plot highlighting mutational patterns in *rpoB* and *rpoC* DNA sequences, with sequences organized according a phylogenetic tree based on a full genome alignment of SNPs.

Figure S2. Positions of *rpoB* and *rpoC* mutations shown in the other figures, noted on drug sensitive consensus protein sequences, to unambiguously highlight their locations within the proteins.

Figure S3. DNA mutations in the *rpoB/rpoC* resistant data set organized by a *rpoB/rpoC* phylogeny, highlighting linked cases.

Supplementary Methods

Table S1. Disease characteristics, treatment regimens, and clinical outcome of subjects whose isolates were used in this study.

Strain identity no.	Previous TB episodes	Initial Chest X-ray	Cavitation	DST	Initial Regimen	Final Regimen	Time to culture conversion (months)	Chest X-ray Change (6mo)	Treatment Outcome
K03b00DS	0	Far advanced	Y	DS	2HERZ/10HER	2HERZ/10HER	4	Improved	Cure
K04b00DS	0	Far advanced	N	DS	2HERZ/10HER	10HER	> 6	Improved	Cure
K05b00DS	0	Minimal	N	DS	2HERZ/7HER	2HERZ/7HER	1	Improved	Cure
K07b00DS	0	Moderately advanced	N	DS	2HERZ/7HER	2HERZ	1	Unchanged	Death
K08b00DS	0	Far advanced	Y	DS	2HERZ/7HER	2HERZ/7HER	1	Improved	Cure
K09b00DS	0	Far advanced	Y	DS	2HERZ/10HER	2HERZ/10HER	1	Improved	Cure
K10b00DS	0	Far advanced	Y	DS	2HERZ/7HER	2HERZ/6HER	2	Improved	Default
K11b00DS	0	Moderately advanced	Y	DS	2HERZ/7HER	2HERZ/7HER	2	Improved	Cure
K12b00DS	0	Moderately advanced	Y	DS	2HERZ/7HER	2HERZ/8HER	2	Improved	Cure
K13b00DS	1	Minimal	Y	DS	2HERZ/10HER	2HERZ/10HER	1	Improved	Cure
K14b00DS	1	Moderately advanced	I	DS	2HERZ/10HER	2HERZ/10HER	1	Improved	Cure
K15b00DS	1	Moderately advanced	Y	DS	2HERZ/10HER	2HERZ/10HER	2	Improved	Cure
K16b00DS	1	Moderately advanced	Y	DS	2HERZ/10HER	2HERZ/13HER	2	Improved	Cure
K17b00DS	2	Far advanced	I	DS	PTCOAmK	2HERZ/12HER	2	Improved	Cure
K18b01PR	1	Far advanced	I	MR	2HERZ/10HER	6ZPTCOS/5ZPTCO	> 2	Improved	Default
K19b00MR	2	Far advanced	N	MR	7PTCLfS/17PTCLf	5ZTCLf	6	Improved	Default
K20b00PR	2	Far advanced	N	PR	7PTCORbK/17PTCORb	10TCORb	4	Improved	Cure
K21b00MR	2	Far advanced	Y	MR	7PTCOS/17PTCO	7PTCLfK/3PTCLf	2	Improved	Default
K22b00MR	2	Far advanced	I	MR	2HERZ/10HER	4ZPTCLfS	4	Improved	Default
K25b00MR	2	Moderately advanced	Y	MR	ERZ	4PTMfAS/17PTMfA	4	Improved	Cure
K26b00MR	2	Moderately advanced	Y	MR	2HERZ/8HER	Unknown	> 1	N/A	Withdrawn
K28b00MR	3	Moderately advanced	Y	MR	7ZTCOK/17ZTCO	4TCO	4	Improved	Death
K29b00MR	6	Far advanced	Y	MR	PTCO	24EZTCLf	4	Improved	Cure
K32b00PR	3	Far advanced	N	PR	7PTCOS/17PTCO	1ZTCLf	2	Unchanged	Default
K32b04PR			N	PR	6ZTCLfS/17ZTCLf				
K33b00PR	0	Far advanced	U	PR	2HERZ/7HER*	16CACILf	> 6	Improved	Death
K33b06XR			Y	XR	7COAClAm/17COACl				
K34u00DR	0	Moderately advanced	N	HR	2HERZ/7HER	2HERZ/6HER	2	Improved	Relapse
K34u13DR			Y	HR					
K35b00DS	2	Far advanced	U	DS	2HERZ/10HER	2HERZ/10HER	6	Improved	Relapse
K35b18MR			Y	MR	N/A				
K36b26PR	6	Moderately advanced	Y	PR	7PTCLfK/17PTCLf	3ZTCMFk/17ZTCMF	1	Improved	Relapse
K37b00XR	5	Far advanced	Y	XR	ECO	2HRZTLS/14HRZTLf	4	Aggravated	Failure

* Regimen changed after 2 months to 7PTCOS/17PTCO

Legend:

Y: Yes

DS: Drug sensitive

N: No

HR: Highly resistant but not formally MR or XR

I: Indeterminate

MR: Multidrug resistant: resistant to isoniazid and rifampicin

U: Unknown

PR: Pre-XDR, resistant to isoniazid and rifampicin and an aminoglycoside or fluoroquinolone but not both

N/A: Not available

XR: Extensively drug resistant, resistant to isoniazid, rifampicin, an injectable and a fluoroquinolone

Drug Key see Table S2 for drug names, regimen is coded by number of months of administration followed by the agents used generally as intensive phase/continuation phase ie. 2HRZE/10HER means 2 months of HRZE followed by ten months of HER in accordance with convention.

Table S2. Drug susceptibility patterns of isolates sequenced in this study.

Strain identity no.	DST by LJ DST kit														Notes.
	INH	RFP	SM	EMB	KM	CPM	PTH	CS	PAS	OFX	MFX	AMK	LEV	RBU	PZA
K03b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	ND	S
K04b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K05b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	ND	S
K07b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K08b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K09b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K10b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K11b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K12b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K13b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K14b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K15b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K16b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K17b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K18b01PR	R	R	S	S	S	S	S	S	S	S	S	S	R	R	S
K19b00MR	R	R	S	R	S	S	R	R	S	S	S	S	R	R	R
K20b00PR	R	R	R	R	R	S	R	S	S	R	S	S	S	S	R
K21b00MR	R	R	R	R	S	ND	S	S	S	ND	ND	ND	ND	ND	S
K22b00MR	R	R	S	R	S	S	S	S	S	S	S	S	R	S	R
K25b00MR	R	R	S	R	S	S	R	S	S	S	S	S	S	R	S
K26b00MR	R	R	R	R	S	S	R	S	S	S	S	S	S	S	S
K28b00MR	R	R	S	R	S	ND	S	S	R	S	ND	ND	ND	ND	S
K29b00MR	R	R	R	S	S	S	S	S	S	S	S	S	R	S	R
K32b00PR	R	R	S	S	R	R	R	S	S	S	R	S	S	S	R
K32b04PR	R	R	S	S	R	R	R	R	S	S	R	S	S	S	R
K33b00PR	R	R	R	R	R	R	R	S	R	S	R	S	R	R	R
K33b06XR	R	R	R	R	R	R	R	S	R	R	R	R	R	R	R
K34u00DR	S	R	S	R	S	ND	S	R	S	S	ND	ND	ND	ND	S
K34u13DR	S	R	S	S	S	R	R	S	R	S	S	S	S	S	S
K35b00DS	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
K35b18MR	R	R	S	R	S	S	S	S	S	S	S	S	S	S	S
K36b26PR	R	R	S	R	S	ND	S	S	S	R	ND	ND	ND	R	S
K37b00XR	R	R	S	R	R	S	S	S	R	R	S	ND	R	R	R

Drug KEY:

INH(H): Isoniazid

1 - DST results were not repeated on the one month isolate from this patient, these results are from the initial isolate.

RFP(R): Rifampicin

SM(S): Streptomycin

2 - DST results and details are for the subjects' initial isolate and disease but the sequence is of the subsequent relapse isolate.

EMB(E): Ethambutol

KM(K): Kanamycin

CPM: Capreomycin

PTH(T): Prothionamide

CS(C): Cycloserine

PAS(P): para-aminosalicylic acid

OFX(O): Ofloxacin

MFX(M): Moxifloxacin

AMK: Amikacin

LEV(L): Levofloxacin

RBU(Rb): Rifabutin

PZA(Z): Pyrazinamide

AUG(A): Augmentin

CLA (Cl): Clarithromycin

2

1

Table S3. Sources of previously published full length Mtb genomic sequences used in this study.

Sequence Designation*	GeneBank ID	Source
US1990_HN878_DS	315064806	Ioerger 2010 ¹
US1993_210_DS	261746034	Filliol 2006 ²
ZAWeC_R1390_IR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_R1441_IR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_R1842_IR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_X122_PR	312987796	Ioerger 2010 ¹
ZAWeC_X132_PR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_X156_PR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_X29_PR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_R1207_MR	312984523	Ioerger 2010 ¹
ZAWeC_R1505_MR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_R1746_MR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_R1909_MR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_X189_XR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_X28_XR	Personal Comm.	Ioerger 2010 ¹
ZAWeC_X85_XR	Personal Comm.	Ioerger 2010 ¹
X_W148_MR	326590567	M. tuberculosis Comparative Database [†]
ZAWeC_F11_IR	148719718	M. tuberculosis Comparative Database [†]
US1905_H37Ra_DS	148503909	Zheng 2008 ³ , Steenken 1946 ⁴ , Kubica 1972 ⁵
US1905_H37Rv_DS	57116681	Cole 1998 ⁶ , Steenken 1946 ⁴ , Kubica 1972 ⁵
ZA1994Dur_1435_MR	253318418	Koenig 2007 ⁷ , M. tuberculosis Comparative Database [†]
ZA1994Dur_V2475_MR	297718568	Ioerger 2009 ⁸
ZA1995Dur_4207_DS	295687135	Koenig 2007 ⁷ , M. tuberculosis Comparative Database [†]
ZA2006Dur_R506_XR	297718569	Ioerger 2009 ⁸
ZA2005TUF_605_XR	289552250	Koenig 2007 ⁷ , M. tuberculosis Comparative Database [†]
CA_SUMu001_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu002_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu003_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu004_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu005_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu006_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu007_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu008_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu009_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu010_DS	NA	TB Natural Mutation Rate Database [‡]
CA_SUMu011_DS	NA	TB Natural Mutation Rate Database [‡]
US_CDC1551_DS	50952454	Fleischmann 2002 ⁹ , van Embden1993 ¹⁰

* Each isolate name is preceded with the ISO (International Organization for Standardization) two-letter country code specifying the country of origin of the sample (United States, US; South Africa, ZA; Canada, CA; and X was used for unknown). This is followed by the year of sampling, if it was specified in the original publication. Samples from South Africa have further geographic resolution specified (WeC for the Western Cape, DUR for Durban, and TUF for Tugela Ferry). This is followed by an underscore and the isolate name as specified in the original publication of the sequence, which is followed by another underscore and the drug sensitivity status of the isolate (drug sensitive, DS; IR isoniazide resistant only; PR, pre-XDR; MR, multidrug resistant; and XR, XDR). [†]Broad Institute, Mycobacterium tuberculosis Comparative Database.

www.broadinstitute.org/annotation/genome/TB_Nature_Mutation_Rate/MultiHome.html. [‡]Broad Institute, TB Natural Mutation Rate Database www.broadinstitute.org/annotation/genome/TB_Nature_Mutation_Rate/MultiHome.html

Table S4. Non-synonymous substitutions in 170 RIF-R clinical isolates.

Isolate	RIF-R mutation in RRDR*	Non-synonymous substitution		
		<i>rpoA</i>	<i>rpoB</i> *	<i>rpoC</i>
1	L511P	-	-	-
2	L511P	-	-	-
3	L511P	-	-	-
4	Q513L	-	-	-
5	Q513E, D516Y	-	N769S	-
6	Q513K	-	R754H	R1085_L1086del
7	Q513L	-	R827C	-
8	M515V, D516G	-	-	-
9	D516L	-	-	-
10	D516N, H526N	-	-	-
11	D516V	-	-	-
12	D516V	-	-	-
13	D516V	-	-	-
14	D516V	-	-	-
15	D516V	-	-	-
16	D516V	-	-	-
17	D516V	-	-	-
18	D516V	-	-	-
19	D516V	-	-	-
20	D516V	-	-	-
21	D516V	-	-	-
22	D516V	-	-	-
23	D516V	-	-	-
24	D516V	-	-	-
25	D516V	-	-	-
26	D516V	-	-	-
27	D516V	-	-	-
28	D516Y	-	-	G594E
29	D516Y	-	D545E, P969S	-
30	D516Y	-	-	-
31	D516Y	-	-	-
32	D516Y	-	-	-
33	D516Y	-	-	-
34	D516Y, N518H	-	D545E	-
35	D516Y, T525I	-	-	-
36	N518del	-	-	A1213E
37	H526C	-	-	-
38	H526C	-	-	-
39	H526C	-	-	-
40	H526D	-	-	A999G
41	H526D	-	H674R	-
42	H526D	-	-	-
43	H526D	-	-	-
44	H526D	-	-	-
45	H526D	-	-	-
46	H526D	-	-	-
47	H526G	-	-	-
48	H526L	-	D897G, K1102T	-
49	H526L	E24K	-	-
50	H526L	-	-	-
51	H526L	-	-	-
52	H526N	-	-	-
53	H526P	-	-	-
54	H526R	-	D92G	-

55	H526S	-	M707V	-
56	H526Y	-	E391G	I565M, H689R
57	H526Y	-	-	E591Q
58	H526Y	-	V570A	-
59	H526Y	-	I1035V	-
60	H526Y	-	-	-
61	H526Y	-	-	-
62	H526Y	-	-	-
63	H526Y	-	-	-
64	H526Y	-	-	-
65	H526Y	-	-	-
66	H526Y	-	-	-
67	S531L	-	-	M143R
68	S531L	-	-	G332R
69	S531L	-	-	G332S
70	S531L	-	-	N416S
71	S531L	-	-	N416S
72	S531L	-	D545E	P434R
73	S531L	-	-	P434V
74	S531L	-	-	K445R
75	S531L	-	-	F452L
76	S531L	-	-	F452L
77	S531L	-	-	F452L
78	S531L	-	-	F452L
79	S531L	-	-	F452L
80	S531L	-	-	F452L
81	S531L	-	-	F452L
82	S531L	-	-	F452L
83	S531L	-	-	F452L
84	S531L	-	-	F452L
85	S531L	-	-	F452L
86	S531L	-	-	F452L
87	S531L	-	-	F452L
88	S531L	-	-	V483A, E49G
89	S531L	-	-	V483G
90	S531L	-	-	V483G
91	S531L	-	-	V483G
92	S531L	-	T1078I	V483G
93	S531L	-	-	W484G
94	S531L	-	-	P495_V496insA
95	S531L	-	-	L507V
96	S531L	-	-	L507V
97	S531L	-	-	L516P
98	S531L	-	-	V517A
99	S531L	-	-	V517L
100	S531L	-	T52P	V517L, E591Q
101	S531L	-	-	L527V
102	S531L	-	-	D714E
103	S531L	-	-	D747A
104	S531L	-	-	H748P
105	S531L	-	-	E750G
106	S531L	-	-	E750G
107	S531L	-	-	E750D
108	S531L	-	Q409R	E750A
109	S531L	-	-	E757A
110	S531L	-	-	E1033A
111	S531L	-	-	P1040R
112	S531L	-	-	P1040S
113	S531L	-	-	A1213V
114	S531L	-	-	N1251S

115	S531L	-	-	V1252L
116	S531L	-	-	V1252L, T853S
117	S531L	-	P45S	-
118	S531L	-	P45S, D545E	-
119	S531L	-	P45S, D545E	-
120	S531L	-	P45S, D545E	-
121	S531L	-	T52P	-
122	S531L	-	F503S	-
123	S531L	-	F503S	-
124	S531L	-	T660P	-
125	S531L	-	H835R	-
126	S531L	A180V	T399I	-
127	S531L	R182W	-	-
128	S531L	V183G	-	-
129	S531L	T187A	D545E	-
130	S531L	D190G	-	-
131	S531L	-	-	-
132	S531L	-	-	-
133	S531L	-	-	-
134	S531L	-	-	-
135	S531L	-	-	-
136	S531L	-	-	-
137	S531L	-	-	-
138	S531L	-	-	-
139	S531L	-	-	-
140	S531L	-	-	-
141	S531L	-	-	-
142	S531L	-	-	-
143	S531L	-	-	-
144	S531L	-	-	-
145	S531L	-	-	-
146	S531L	-	-	-
147	S531L	-	-	-
148	S531L	-	-	-
149	S531L	-	-	-
150	S531L	-	-	-
151	S531L	-	-	-
152	S531L	-	-	-
153	S531L	-	-	-
154	S531L	-	-	-
155	S531L	-	-	-
156	S531L	-	-	-
157	S531L	-	-	-
158	S531L, D516Y	S82R	-	K1256E
159	S531Q	-	-	-
160	S531W	-	-	K715T
161	S531W	-	-	T812I
162	S531W, S512R	-	-	-
163	L533P	-	-	M143R
164	L533P	-	-	-
165	L533P	-	-	-
166	L533P	-	-	-
167	L533P	-	-	-
168	L533P	-	-	-
169	L533P	-	-	-
170	L533P	-	-	-

* According to convention the numbering of the *E. coli* enzyme is used when referring to mutations within the RRDR of *rpoB* and according to the *M. tuberculosis* numbering when referring to mutations outside of that region of *rpoB*.

Table S5. MIRU-VNTR patterns of selected isolates showing convergent acquisition of compensatory alleles for *rpoB* S531L mutation and transmission of compensated strains.

Allele	Isolate	Locus																							
		154	424	577	580	802	960	1644	1955	2059	2163b	2165	2347	2401	2461	2531	2687	2996	3007	3171	3192	3690	4052	4156	4348
<i>Transmission of compensated strains</i>																									
<i>rpoC F452L</i>	75	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	76	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	77	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	78	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	79	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	80	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	81	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	82	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	83	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	84	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	85	2	2	4	2	3	3	3	6	2	5	4	2	4	2	5	1	7	3	3	5	4	8	3	2
	86	2	2	4	2	3	3	3	6	2	5	3,4	2	4	2	5	1	7	3	3	5	4	8	3	2
	87	2	2	4	2	3	3	3	6	2	5	3	2	4	2	5	1	7	3	3	5	4	8	3	2
<i>rpoB P45S</i>	118 [†]	2	3	4	2	3	3	4	4	4	7	1,4	4	4	2	5	7	3	5	3	2	4	3		
	119 [†]	2	3	4	2	3	3	4	4	4	2	7	4	4	2	5	1	7	3	3	5	3	2	4	3
	120 [†]	2	3	4	2	3	3	4	4	4	2	7	4	4	2	5	1	7	3	3	5	3	2	4	3
<i>rpoB L507V</i>	117 [*]	2	3	4	2	3	2	3	4	2	4	4	4	2	5	1	7	3	3	5	3	8	3	3	
	95	2	3	4	2	2	3	3	5	2	5	4	4	4	2	5	1	7	3	3	4	4	8	3	3
<i>rpoC E750G</i>	96	2	3	4	2	2	3	3	5	2	5	4	4	4	2	5	1	7	3	3	4	4	8	3	3
	105	2	2	4	2	3	3	3	4	2	4	4	4	2	5	1	7	3	3	5	4	7	3	3	
	106	2	2	4	2	3	3	3	4	2	4	4	4	2	5	1	7	3	3	5	4	7	3	3	
<i>Convergent acquisition of compensatory alleles</i>																									
<i>rpoC V483G</i>	89	2	4	4	2	3	3	3	5	2	6	4	4	4	2	5	1	9	3	3	4	5	10	2	3
	90	2	2	4	2	3	3	3	4	2	6	2	4	4	2	5	1	7	3	3	5	2	7	3	3
	91	2	4	4	2	3	3	3	4	2	6	4	4	4	2	5	1	7	4	3	5	3	8	4	2
	92 [‡]	2	2	4	2	3	3	3	8	2	5,6	4	4	4	2	5	1	7	3	3	5	3	8	3	3
<i>rpoC N416S</i>	70	2	3	4	2	2	3	3	5,8	2	5	4	4	4	2	5	1	7	3	3	6	3	9	3	3
	71	2	2	6	2	3	3	3	4	2	6	4	2	4	1	5	1	7	3	3	5	4	8	1	3

<i>rpoB</i> F503S	122	2	2	4	2	3	3	3	4	3	6	4	4	4	2	2	1	8	3	3	5	3	8	3	3
	123	2	2	4	2	4	3	3	4	2	5	4	4	4	2	5	1	5	3	3	5	3	8	3	4
<i>rpoB</i> T52P	121	2	3	4	2	2	3	3	5	2	5	4	4	4	2	5	1	7	3	3	4	3	9	3	3
	100 [§]	2	4	4	2	3	3	3	4	2	3	4	4	4	2	2	5	1	7	3	3	2	3	7,8	3
<i>rpoC</i> V1252L	115	2	3	4	2	2	3	3	5	2	5	4	4	4	2	5	1	7	3	3	3	3	9	3	3
	116 ^{§§}	2	3	4	2	2	3	3	5	2	5	4	4	4	2	5	1	7	3	3	4	3	9	3	3

Additional alleles: [†]*rpoB* D545E, [‡]*rpoB* 1078I, [§]*rpoC* V517L *rpoB* E591Q, ^{§§}*rpoC* T853S

* This isolate seems to have evolved separately from other three isolates, not being a part of the transmission chain.

Table S6. Bacterial strains and plasmids.

Strain/plasmid	Description
<i>Plasmids</i>	
pRH1351	a derivative of pPR23: <i>ori</i> ^{ts} (pAL5000): <i>sacB xyIE</i>
pTS421	a derivative of pMV306 (MedImmune) carrying hygromycin-resistance gene
pTS422	pTS421 <i>M. tuberculosis rpoB rpoC</i>
pTS423	pTS421 <i>M. tuberculosis rpoB C1349T rpoC</i>
pTS424	pTS421 <i>M. tuberculosis rpoB C1349T rpoC T1354C</i>
pTS428	pRH1351 carrying the 5' flanking sequence of <i>rpoB</i> (upstream 923-bp plus 45-bp at 5' end of <i>rpoB</i>) and the 3' flanking sequence of <i>rpoC</i> (21-bp at 3' end of <i>rpoC</i> plus downstream 1,003-bp) of <i>M. smegmatis</i> joined with zeocin-resistance marker
pTS429	pTS421 <i>M. tuberculosis rpoB C1349T rpoC T1448G</i>
<i>Strains</i>	
TS102	<i>M. smegmatis mc</i> ² 155 <i>attB::pTS422</i>
TS103	<i>M. smegmatis mc</i> ² 155 <i>attB::pTS423</i>
TS105	<i>M. smegmatis mc</i> ² 155 <i>attB::pTS424</i>
TS106	TS102 <i>ΔrpoB rpoC</i> _{<i>M. smegmatis</i>}
TS108	TS103 <i>ΔrpoB rpoC</i> _{<i>M. smegmatis</i>}
TS110	TS105 <i>ΔrpoB rpoC</i> _{<i>M. smegmatis</i>}
TS112	<i>M. smegmatis mc</i> ² 155 <i>attB::pTS429</i>
TS113	TS112 <i>ΔrpoB rpoC</i> _{<i>M. smegmatis</i>}

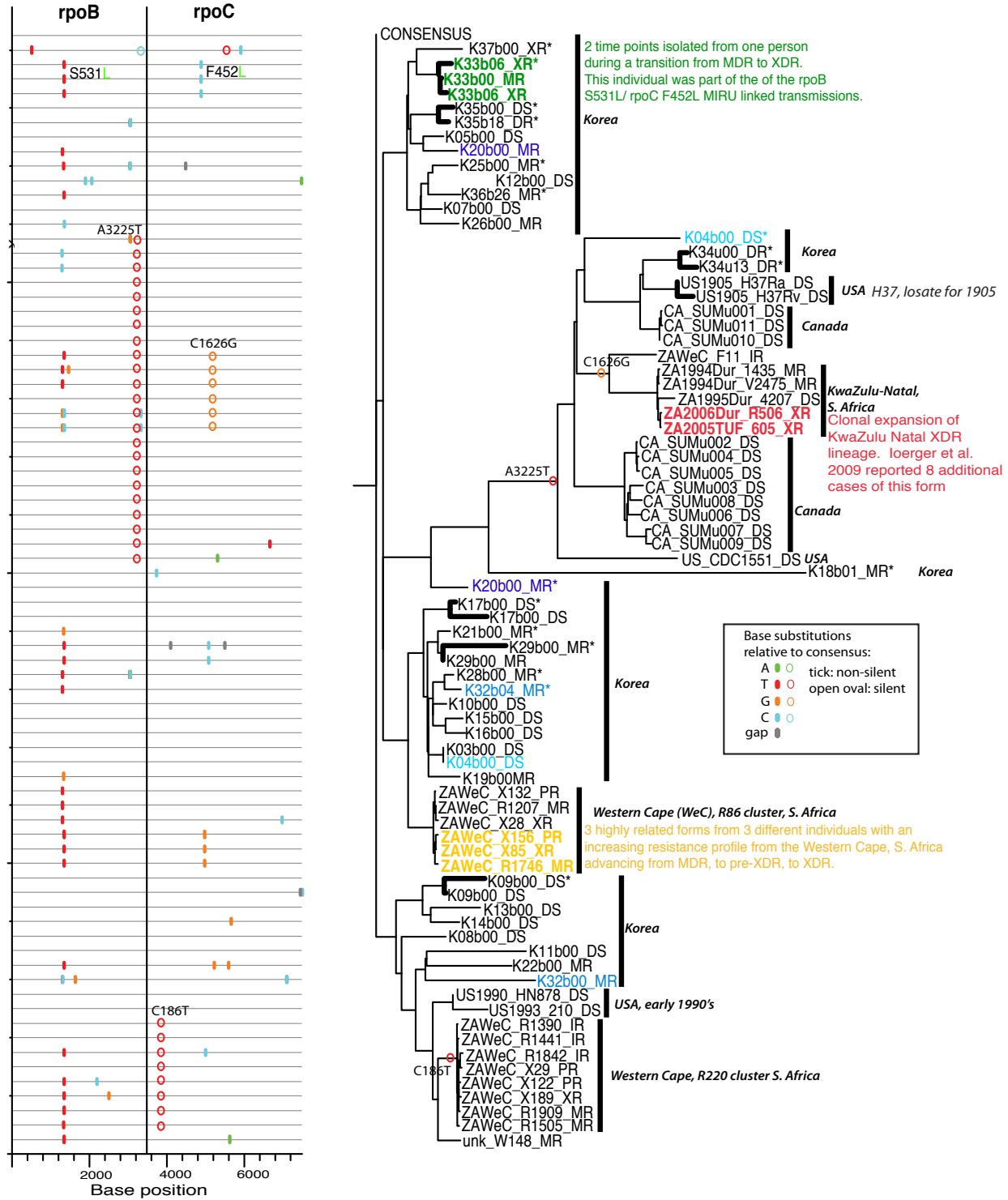


Figure S1. A plot highlighting mutational patterns in *rpoB* and *rpoC* DNA sequences, with sequences organized according a phylogenetic tree based on a full genome alignment of SNPs. In Fig. 2 in the main body of the paper, RpoB and RpoC protein mutations are shown organized according to drug resistance mutations in RpoB. In contrast, here we show the *rpoB* and *rpoC* DNA mutations highlighted relative to the consensus form of the sensitive strains, and ordered to correspond with the phylogenetic tree on the right. Sequence nomenclature is

described in Fig. 2. Several interesting relationships are highlighted. First, only one subject included in the full-length genome set had the combined *rpoB* S531L *rpoC* F452L mutations that appear to be part of a transmitted lineage in Korea. This subject was sampled during the transition from MDR to XDR TB, and sequenced in 2 different laboratories, and the paired mutations are maintained through the period of acquisition of XDR TB and evident in all three sequences (marked in green). Second, the pair marked in red are part of a clonal expansion spreading within the XDR outbreak in KwaZulu-Natal; 8 additional isolates highly related to this form were also identified⁸ when the sequences of these 2 representatives were obtained. Third, the gold group marks another previously sequenced set identified in 2010 from the Western Cape in S. Africa¹. These isolates were originally thought to not represent transmission of drug resistant forms, because not all drug resistant mutations were shared among these subjects when considering multiple genes and drugs. However, the mutation in *rpoC* may be compensatory, and this combined with the extremely close relationship of the isolates in the SNP tree suggests another scenario that is consistent with the data: A RIF resistant form with improved fitness due to a compensatory *rpoC* mutation may have been transmitted, and different levels of multidrug resistance acquired building onto the initial RIF resistance. Fourth, synonymous substitutions (marked by open circles in the plot on the left) were uncommon, and the three recurrent synonymous mutations tracked perfectly with distinct genetic lineages, indicated in the presumed ancestral node in the tree on the right. Finally, bold branches in the tree indicate highly related samples from the same individual that were sequenced more than once, in two different reference laboratories. In 5 cases sequences from the same individual were genetically highly related. In the other 2 cases (marked in shades of blue) the sequences were not related, raising the possibility that these two individuals were infected by more than one strain.

A) RpoB

126 133
LADSRQSKTAASPSPSRPQSSNNSV**P**GAPNRV**S**FAKLREPLEVPGLLDVQTDSFEWLIGSPRWRESAAE
S P
RGDVNPVGGLEEVLYELSPIEDFSGSMSLSFSDPRFDDVKAPVDECKDKDMTYAAPLFVTAEFINNNTGE
IKSQTVFMGDFPMMTEKGTFIINGTERVVSQLVRSPGVYFDETIDKSTDKTLHSVKVIPSRGAWLEFDV
DKRDTVGVRIDRKRRQPVTLLKALGWTSEQIVERFGFSEIMRSTLEKDNTVGTDEALLDIYRKLRPGE**P**
PTKESAQTLLENLFFKEKRYDLARVGRYKVNKKLGLHVGEPISTSSTLEEDVVATIEYLVRLHEGQTTMT
472 480 490
VPGGVEVPVETDDIDHFGNRLRTVGELIQNQIRVGMSRME**R**VVRERMT**T**TQDVEAITP**Q**TLINIRPVVA
G I R
511 516 526 531 554 568
IKEFFGTSQ**L****S****Q****F****M****D****Q****N****N****P****L****S****G****L****T****H****K****R****R****L****S****A****I****L****G****P****G****G****L****S****R****E****R****A****G****L****E****V****R****D****V****H****P****S****H****Y****G****R****M****C****P****I****E****T****P****E****G****P****N****I****G**
PRL VV H- ID L P P S
K Y Y W
E G L Q
N P
R
G
C
S
N
584 626
IGSLSVYARVNPF**G**FIETPYRKVVDGVVSDEIVYLTADEEDRHVVAQANSPIAD**D**GRFVEPRVLVRRKAG
S E
651
EVEYVPSSE**V**DYMDVSPQMVSVATAMIPFLEHDDANRALMGANMQRQAVPLVRSEAPLVGTGMELRAAI
A
741 755
DAGDVVVAEESGVIEEVSA**D**YITVMHDNG**T**RRTYMRKFARSNH**G**TCANQCPIVDAGDRVEAGQVIADGP
P R
788 812 835
CTDDGE**M**ALGKNLLVAIMPWEGHNYEDAI**I**LSNR¹VEEDVLTSIHIEEH**I**DARTKLGAE**E**ITRDIPNI
G P H
916
SDEVLADLDERGIVRIGAEVRGDILVGKVTPKG**E**LTPEERLLRAIFGEKAREVRDTSLKVP**H**GESGK
R
978
VIGIRVFSREDEDEL**P**AGVNELRVYVAQKRKISDGDKLAGRHGNKG**V**IGKILP**V**E**D**MPFLADGTPV**D**
G
LNTHGVPRRMNIGQILETHLGWCAHSGWKVDAAKGVPDWAARLPDELLEAQPNIAVSTPVFDGAQE**A**
ELQ 1116
GLSCTLPNRGDVLVDADGKAMLFDGRSGEPFPYPVTGYMYIMKLHHLVDDK**I**HARSTGPYS**M**ITQQP
V
1159 1177 1183 1187
LGGKAQFGGQRF**G**EMECWAMQAYGAAY**T**LQELLTIKSDDTVGRVK**V**YEAV**K**GEN**I**PEPGIPESFKVLLK
I G T T
ELQSLCLNVEVLSSDGAAIELREGEDDLERAANLGINLSRNESASVEDLA

B) RpoC

VLDVNFFDELIGLATAEDIHQWSYGEVKKPETINYRTLKPEKDGFLCEKIFGPTRDWECYCGKYKRVRF
KGIICERCGVEVTRAKVRERMGHIELAAPVTHIWFKGVPNSRLGYLLDIAPKDLEKIIYFAAYVITSVD
EEMRHNLSTLEAEMAVERKAVEDQRDGELEARAQLEADLAELAEAGAKADARRKVARDGGEREMRQIRD
RAQRELDRLEDIWSTFTKLAPQQLIVDENLYRELVDRYGEYFTGAMGAESIQKLIENFDIDAEAESLRDV
332
IRNGKGQKKLRALKLVAAFQQSGNSPMGMVLDAPVPIPPELRPMVQLDGGRFATSDLNDLYRRVINR
R
S
NNRLKRLIDLGAPIIVNNEKRLQESVDALFDNGRRGRPVTPGNRPLKSLSDLLKGKQGRFRQNLLGK
434 452 483
RVDYSGRSVIVVG**P**QLKLHQCGLPKLMALLE**F**KPFVMKRLVDLNHAQNIKSAKMVERQRPQ**V**WDVLEEV
V L GG
R A
492 507 517 528
IAEHPVLLNRAPTLHRL**G**IQAFEPMLVEGKAIQLHPL**V**CEAFNADFQGDQMAVHLPLSAEAQAEARILML
T V P V
565 591 594
SSNN**I**LSPASGRPLAMPRLD**M**V**T**GLYYLTTE**E**VP**G**D**T**GEY**Q**PASGDHPETGVYSSPAEAIMAADRGVLSVR
M G Q E
689 698
AKIKVRLTQLRPPVEIEAELFGHSGWQPGDAWMAETTLGRVMFNELLPLGYPFVNQ**M**KKVQAAI**I**N**D**
R R
711 714 748 757
AERYPMIVVA**Q**TV**D**KLKDAGFYWATRSGVTVSADVLPVPRKKE**I****D****H****Y**ERAD**K****V**E**K****Q****F**QRGALNHDER
R ET AP G A
A
812
NEALVEIWKEATDEVGQALREHYPDDNPIITIVD**S**GATGNFT**T**QTRTLAGMKGLVTNP**K**GEFIPR**P**V**K**SSF
I
853
REGLT**V**LEYFINT**H**GARKGLADTA**R**TADSGYL**T**RR**L**VD**V**S**Q**D**V**IREHDC**Q**TERGIVVELAERAPDG**T**
S
IRDPYIETSAYARTL**G**TD**A**V**D**EAGNV**V**ERG**Q**DL**G**PE**I**D**A**LLAAG**I**TO**Q**V**K**VR**S**VL**T**C**A**T**S****T****G****V**C**A**T**C****G**
994 1033 1040 1044
RSMATGKLVD**I**GE**A**VG**I**VAAQS**I**GE**P**GT**Q**LT**M**RT**F**H**Q**GG**V**GED**I**T**G**GL**P**RV**Q****E**LF**E**AR**V****P**RG**K****A**PI**A**D**V**
G A S V
R
GRVRLEDGERFYKITIVPDDGEEVVYDKISKRQRLRVFKHEDGSERVLS**D**GD**H**VEVG**Q**QLMEGSADP**H**
1150 1190
VLRV**Q**GPREV**Q**IHLV**R**EV**Q**EV**Y**RA**Q**GV**S****I****H**DKHIEV**V**R**Q**ML**R**VT**I****I**D**S****G****S****T****E****F****L****P****G****S****L****I****D****R****A****E****F****E****A****E**
P N
1195 1216 1255 1259
NRRVV**-**AEG**G**EP**A**AGR**P**V**L**MG**I**T**K****A****S****L****A****T****D****S****W****L****S****A****A****S****F****Q****E****T****T****R****V****L****D****A****A****I****N****C****R****S****D****K****L****N****G****L****K****E****N****V****I****I****G****K****L**
A E SL E
V
I PAGTG**I**NRYRNIAV**Q**PTEEARAAAY**T**IPS**Y**ED**Q**YY**S**PDFGAAT**G**AAV**P**L**D**GYSD**Y**R

Figure S2. Positions of *rpoB* and *rpoC* mutations shown in the other figures, noted on drug sensitive consensus protein sequences, to unambiguously highlight their locations within the proteins.

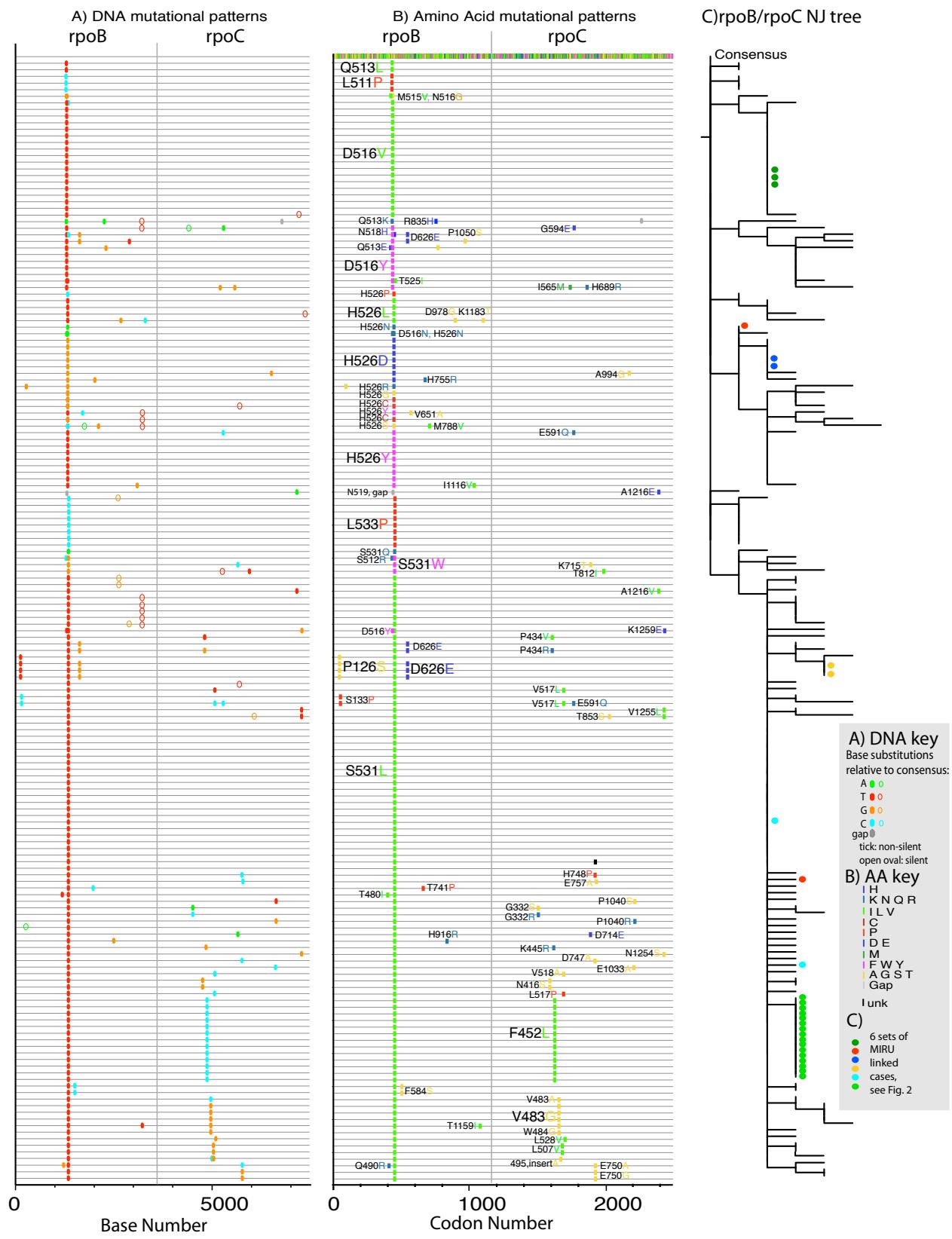


Figure S3: DNA mutations in the *rpoB/rpoC* resistant data set organized by a *rpoB/rpoC* phylogeny, highlighting linked cases. (A) DNA mutations in *rpoB/rpoC* found in the 170 isolates studied. (B) The corresponding amino acid substitutions for these isolates. (C) A Neighbor-joining phylogenetic tree based just on the sequences from these two genes, so that similar mutational patterns in these genes would be naturally grouped. The sequences in (A) and (B) are organized by this tree. Isolates with nearly identical MIRU-VNTR patterns and so very closely related (data taken from the MIRU-VNTR patterns in Fig. 3B in the paper) are highlighted. The dark green, light green, dark blue and gold dots may represent transmission of RIF resistant forms. The light blue and red dots are highly related isolates, but carry distinct patterns of mutation for RIF resistance.

Supplementary Methods

Creating Alignments from Full-length Genomic sequences. We began by using Velvet¹¹, a *de novo* genomic assembler, to assemble raw Illumina-Solexa data into contigs. Contigs that had less than 200 bases were discarded. For each contig, a BLASTN search was performed against the W148 reference genomic sequence, its W149 location identified, and then contigs were arranged along the W148 reference strain according to their locations. When an inconsistency was found in two or more overlapping contigs, we used the base call in the longer stretch. The data were assembled into partial chromosome sequences (most were not complete), and we used BLASTN in conjunction with code we developed to group homologous sequences based on the 3,933 genes identified in the W148 reference strain, into gene files. Each gene file has no more than one nucleotide sequence from each isolate, but not all isolates have sequences that spanned all genes. The criteria we used to call homologous gene sequences was to select the best match, requiring a minimum identity of 90%, and also requiring that the sequence spanned at least 75% of the reference gene. The homologous sequences were then codon aligned, using a method that incorporated the MUSCLE alignment tool¹². To minimize inclusion of localized sequencing errors in these artificial SNP sequences, we first excluded sequences with SNPs that clustered in a local region in a single isolate's gene sequence, specifically where three SNPs were found in a 10 base pair window when compared to reference genes. An artificial SNP sequence was constructed by concatenating all positions where a SNP was identified (where at least one sequence differed from other sequences) for phylogenetic analysis.

Sequence Analysis. Distance-based Neighbor-joining trees were generated from either SNP alignments or concatenated *rpoB* and *rpoC* genes using the PAUP program as implemented in the Los Alamos HIV Database

(<http://www.hiv.lanl.gov/components/sequence/HIV/treemaker/treemaker.html>); plots that highlight mutational patterns were created using the Los Alamos HIV database Highlighter tool (http://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html). To generate the data regarding the relative frequency of mutations among drug-sensitive and drug-resistant strains used in Fig. 1, we first calculated the fraction of non-synonymous SNPs by dividing number of non-synonymous SNPs by the total number of bases for each gene alignment among either the drug-sensitive or resistant sequences. To avoid division by zero when subsequently taking the ratio of these fractions, we used a Laplace estimate of the frequencies (i.e. if a is the number of non-synonymous base substitutions, and n the total number of bases, we estimated the fraction as $(a+1)/(n+2)$). We then plotted the fraction of non-synonymous substitutions in the resistant strains divided by the fraction of non-synonymous substitutions in the sensitive strains on the y axis, by the total number of SNPs found in the gene alignment, in Fig. 1. Gene alignments were excluded from the analysis if they appeared to be heavily mutated (more than 5 changes on average per gene), if they did not have a minimum of 3 sensitive and 3 resistant strains, if they were a member of a heavily repeated multigene families PPE and PE-PGRS (which tend to have problematic alignments), or if they were hypothetical coding regions. This left 2,558 genes for inclusion in the analysis.

References

1. Ioerger, T. R., Feng, Y., Chen, X., Dobos, K. M., Victor, T. C., Streicher, E. M., et al. (2010) The non-clonality of drug resistance in Beijing-genotype isolates of *Mycobacterium tuberculosis* from the Western Cape of South Africa. *BMC Genomics* **11**: 670.
2. Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbón, M. H., Bobadilla del Valle, M., et al. (2006) Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* **188**: 759-772.
3. Zheng, H., Lu, L., Wang, B., Pu, S., Zhang, X., Zhu, G., et al. (2008) Genetic basis of virulence attenuation revealed by comparative genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv. *PLoS One* **3**: e2375.
4. Steenken, W., Jr., and Gardner, L. U. (1946) History of H37 strain of tubercle bacillus. *Am Rev Tuberc* **54**: 62-66.
5. Kubica, G. P., Kim, T. H., and Dunbar, F. P. (1972) Designation of strain H37Rv as the neotype of *Mycobacterium tuberculosis*. *Int J Systematic Bacteriology* **22**: 99-106.
6. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537-544.
7. Koenig, R. (2007) Tuberculosis. Few mutations divide some drug-resistant TB strains. *Science* **318**: 901-902.
8. Ioerger, T. R., Koo, S., No, E. G., Chen, X., Larsen, M. H., Jacobs, W. R. Jr., et al. (2009) Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. *PLoS One* **4**: e7778.
9. Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., et al. (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* **184**: 5479-5490.
10. van Embden, J. D., Cave, M. D., Crawford, J. T., Dale, J. W., Eisennach, K. D., Gicquel, B., et al. (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* **31**: 406-409.
11. Zerbino, D. R., Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821-829.
12. Edgar, R. C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.